

Data & Emoce

ANALYTICKÁ DÍLNA DATA FESTIVALU

Filip Vencovský
Katedra informačních technologií

filip.vencovsky@vse.cz

Cíle analytické dílny

- Diskutovat možnosti a omezení analýzy emocí z textu pro marketingové využití
- Prakticky si vyzkoušet analýzu sentimentu

Technické požadavky

- Analytická platforma KNIME
- Balíček pro textovou analytiku
- Zdrojová data

- Popis požadavků a data

<https://github.com/filipvencovsky/datafestival>

Analytická platforma KNIME

- Stažení z webu

<https://www.knime.com/downloads/download-knime>

- Instalace balíčku pro textovou analytiku
 - Help > **Install New Software**
 - Work with: --All Available Sites—
 - Hledat klíčové slovo „text“
 - Označit **KNIME Textprocessing 3.5.3**
 - Nainstalovat podle průvodce

Sebevražední básníci

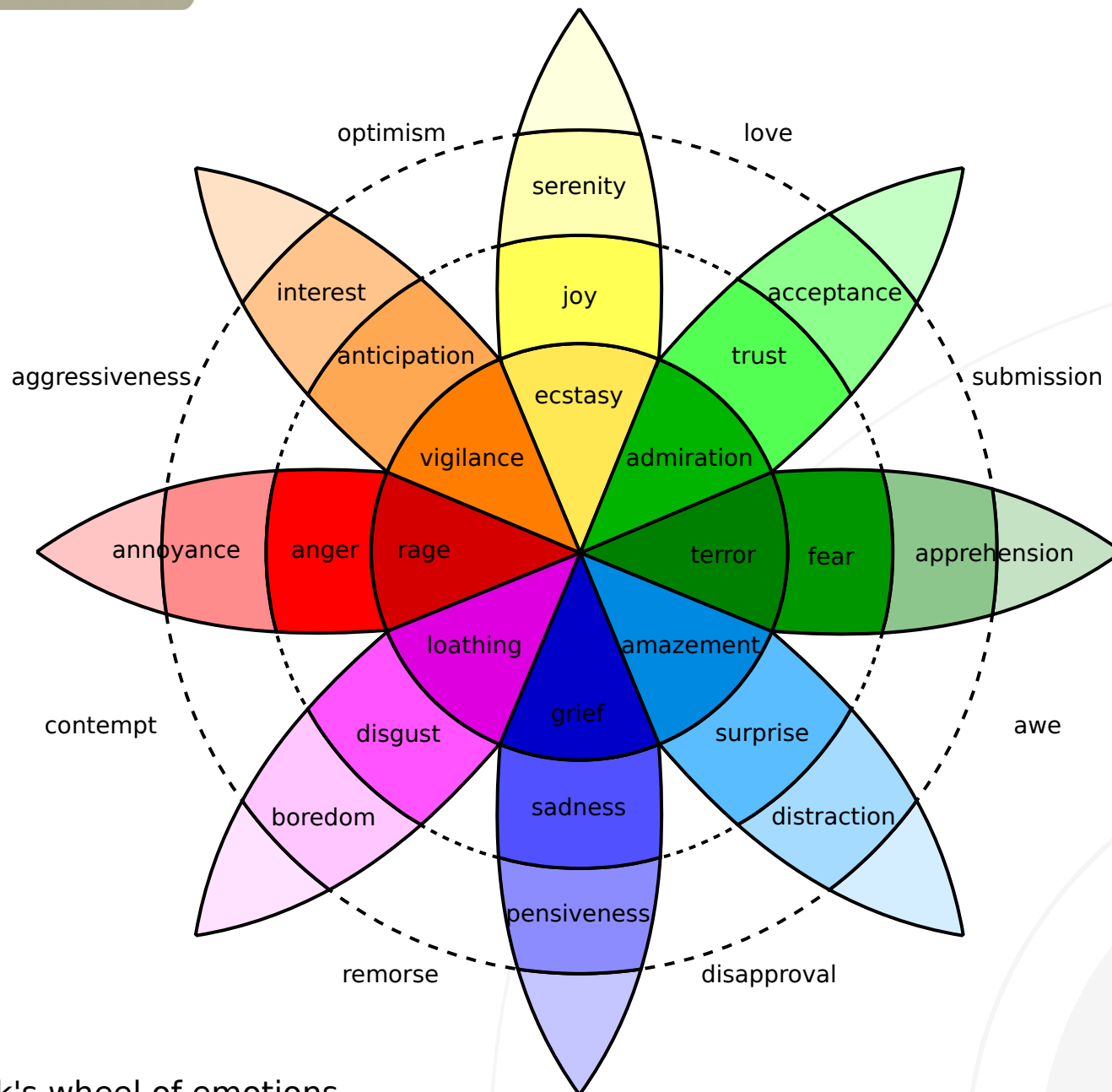
- Wiltsey Stirman, Shannon MA; Pennebaker, James W. PhD. **Word Use in the Poetry of Suicidal and Nonsuicidal Poets.** Psychosomatic Medicine: July-August 2001 - Volume 63 - Issue 4 - p 517-522
- Strojová analýza textu
- 300 básní od 9 básníků, kteří spáchali a 9, kteří nespáchali sebevraždu
- Významná odlišnost mezi skupinami v použití první osoby v jednotném a množném čísle
- Možnost určení rizika sebevraždy na základě analýzy textu

Využití analýzy emocí z textu?

Emoce

- Téma bohaté na definice v různých oborech
- NAKONEČNÝ, M. 1997. Encyklopedie obecné psychologie. Praha: Academia. ISBN 80-200-0625-7
 - „**mentální stav** charakterizovaný cítěním s různým tělesným doprovodem, který vyjadřuje **vztah k objektu**.”
 - Dále jsou emoce excitovaný stav mysli, **doprovázený** cíleně zaměřeným **chováním**, afektivním stavem, který byl vyvolán určitou překážkou nebo dočasným nedokončením akce.
 - Jde o dynamický projev instinktu, který následuje jako organizovaná **odpověď organismu na podnět** a to převážně prostřednictvím autonomní nervové soustavy.”
- Reakce na nerovnováhu, směřování zpět do nového bodu

Jaké jsou druhy emocí?



Plutchik's wheel of emotions

Modely klasifikace emocí

- Obvykle kruhové uspořádání
 - kromě Plutchika, také Scherer či Russel
 - Ukazuje jak intenzitu, tak spojité přechody
- Modální emoce jsou zobecnění problému
 - radost, důvěra, strach, překvapení, smutek...
- Sentiment je potom ještě vyšší forma zobecnění
 - jen pozitivní a negativní ladění
 - obvykle také intenzita

Modelová situace

- Hotelový pokoj bez objednaného výhledu
 - Jaké emoce může člověk pociťovat?
 - Co je předmětem emocí?
 - Budou emoce vidět? Jak?
 - Jaké chování situace podněcuje?

„Měření“ emocí

- Ideální, uvnitř
 - mozek, nervová soustava, hladina hormonů
 - možné velmi omezeně
- z fyzických projevů, navenek
 - gesta, výraz tváře
 - běžné v komunikaci, technicky možné
 - Ve správný čas, ale pomíjivé
- „self-reported“, zprostředkované
 - dotazování, zanesení bodu na kruh
 - Možné i později, ale může být zkreslené

Jaká data mají firmy o zákaznících k dispozici?

Modelová situace II

- Přijedete domů a rozhodnete se napsat recenzi
 - Kolik si toho budete pamatovat?
 - Co vše mohlo vzpomínku ovlivnit?
 - Jak moc se emoce do textu vtisknou?

Emoce a paměť

- Úzká souvislost
- V mozku
 - Hipokampus slouží k uchování fakt a událostí
 - Amygdala k jejich citovému zabarvení
- Emočně významnější události se lépe zapamatují
 - Tvar písmene U potažmo J u online recenzí

Omezení analýzy textu I

- Text není projev emocí, ale jejich reflexe
- Do textu se dostanou až s časovou prodlevou
- Závisí na schopnosti psát text

Opinion mining

- Obor, zkoumající názory vyjádřené v textu
- Bing Liu. 2015. Sentiment Analysis: Mining Opinions, Sentiments and Emotions
- Rozlišuje se:
 - Držitel názoru
 - Cílová entita a její aspekty
 - Sentiment a jeho intenzita, případně jiné emoce
 - čas zachycení názoru

Obsah vs. forma

- Zjednodušeně jazyk obsahuje
 - Výrazy nesoucí význam – **obsahová**,
např. stůl, nést, hluboký, přehledně
 - Výrazy, která obsah propojuje – **funkční**,
např. to, s, !, ačkoli
- Kolik je kterých?
- Jak často se v jazyce vyskytují?

Obsahová slova

- Kromě významu mají i emoční zabarvení
 - Auto, vůz, rachotina, žihadlo
- Záleží na kontextu užití
 - žihadlo – rychlé auto, orgán včely
- Sdružují se do frází, které mohou emoční zabarvení výrazně měnit
 - příležitost x promarněná příležitost

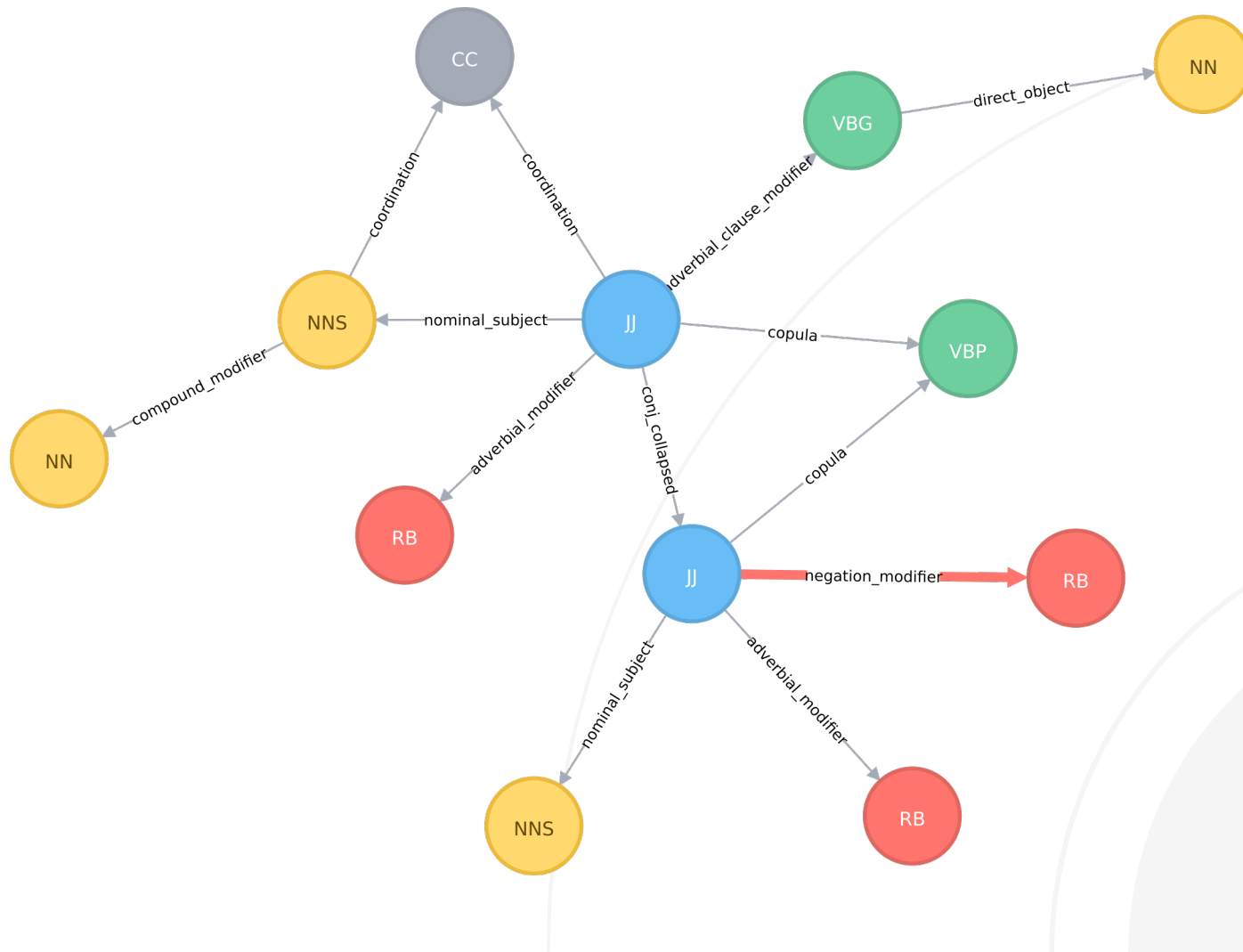
Funkční slova a styl

- Specifické použití funkčních slov je projevem stylu autora a odráží
 - mentální stav jedince
 - ale také sociální status
- Pennebaker, JW. 2011. The Secret Life of Pronouns. ISBN 9781608194971
- Hodí se spíše na zkoumání osobnosti, sociálního začlenění (viz básníci)

Závislosti mezi slovy

- Každé slovo má vazbu nejméně na jedno další
- Strojové zpracování je náročné na výkon
- Popisují se formou stromu
 - Např. projekt Universal Dependencies
- Z hlediska emocí jsou nejzajímavější **modifikátory**
 - promarněná příležitost, strašně zajímavé

Závislosti mezi slovy



Omezení analýzy textu II

- Emoční zabarvení se mění
 - S dalšími významy slova
 - Spojením s jinými slovy
 - Užitím jazyka v různém kontextu, např. skupiny, oboru

Možnosti klasifikace z textu

- Slovníky
 - NRC, LIWC, WordNet-Affect
 - Dosahují horších výsledků při klasifikaci
 - Obsahují více slov
- Strojové učení s učitelem
 - Otázka složení vstupních dat pro učení
 - Slova, všechna? která?
 - Skupiny po sobě jdoucích slov (n-gramy)
 - Páry slov podle vzájemného vztahu
 - Co použít jako váhy?
 - Výskyt slova, relativní četnost slova
 - Významnost slova v datasetu a dokumentu

Příprava textu pro analýzu

- Rozdělení textu na věty a slova
- Obohacení o slovní druhy, vztahy mezi slovy, rozpoznání vlastní jména
- Sjednocení tvarů slov
 - Kvůli hledání ve slovnících a zjednodušení modelu
 - Zkrácení na kořen – stem, nebo převedení na základní tvar slova – lemma
 - Sjednocení velikosti písmen
- Filtrování slov
 - Příliš četná, málo četná
 - Někdy také interpunkce, stop slova, číslovky
 - Pro zjednodušení modelu

Praktická část

Analýza sentimentu

Zdrojová data

- 3 tisíce vět v angličtině
- Yelp, Amazon, IMDB
- Asociace se sentimentem
- Obohacené o klasifikaci pomocí Stanford CoreNLP
- Ke stažení v repozitáři na Github.com

<https://github.com/filipvencovsky/datafestival>

Práce s analytickou platformou...

Děkuji za pozornost

Dotazy?

Filip Vencovský
Katedra informačních technologií

filip.vencovsky@vse.cz